How to Evaluate the Usefulness and Quality of Synthetic Data

Jeremy Wu, Assistant Division Chief LEHD, Data Integration Division
John M. Abowd, Distinguished Senior Research Fellow

Census Advisory Committee of Professional Associations
October 26-27, 2006

Abstract

The use of synthetic data is emerging as a viable approach to making analytically valid data available for public use while strictly protecting the confidentiality of the respondents to the original data. For example, the Census Bureau released the first partial synthetic data product known as On The Map—a web-based mapping application showing where people live and work—under the Longitudinal Employer-Household Dynamics (LEHD) program in February 2006. Experimental development has also started with the re-engineering of the Survey of Income and Program Participation (SIPP) known as Dynamics of Economic Well-being System (DEWS). While Bayesian techniques and integrated data provided the theoretical foundation and the estimation of their respective posterior predictive distributions, evaluation of the usefulness and quality of synthetic data is an ongoing process. The authors will present theoretical work that has been developed in a general setting and applied to On The Map and DEWS. Ongoing activities will also be described, including the National Science Foundation grant on synthetic data development and practical improvement to their usefulness and quality.

Questions to Committee

1. How can professional organizations help to develop and promote standards for integrated data and synthetic data to understand what they are and evaluate their usefulness and quality?

2. How should the Census Bureau communicate the standards for synthetic data quality to the American public?

This document reports the results of research and analysis undertaken by the U.S. Census Bureau staff. This document is released to inform interested parties of ongoing research and to encourage discussion of work in progress. This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD), which is partially supported by the National Science Foundation Grants SES-9978093 and SES-0427889 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging Grant R01 AG018854, and the Alfred P. Sloan Foundation. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau, its program sponsors or data providers. Some or all of the data used in this paper are confidential data from the LEHD Program. The U.S. Census Bureau supports external researchers' use of these data through the Research Data Centers (see www.ces.census.gov). For other questions regarding the data, please contact Jeremy S. Wu, ADC, U.S. Census Bureau, LEHD Program, Data Integration Division, FOB 3, Room 2138, 4700 Silver Hill Rd., Suitland, MD 20233, USA. (Jeremy.S.Wu@census.gov http://lehd.dsd.census.gov).

INTRODUCTION

Responding to a report (1993) by the Panel on the Confidentiality and Data Access, Rubin (1993) proposed a new research initiative to release "only synthetic microdata sets" for general public use while retaining the analytical validity of the original data. In the same volume, Little (1993) proposed the approach that has become known as "partially synthetic" microdata, where only the "sensitive values" are replaced by synthesized data. Traditional methods (coarsening, top coding, swapping, and cell suppression) deny users access to certain data, which have been masked to protect the confidentiality of the original provider. Businesses and individuals, who supplied the original data either as direct respondents or as the source of administrative records, are entitled to this confidentiality protection according to both legal and ethical standards. The foundation of the American official statistical system is predicated on the trust that citizens place on the stewards of the data to uphold these standards.

While there are now a variety of approaches to Rubin's original proposal, the term "fully synthetic data" is now understood within the statistics community to mean that the original confidential data have been used to simulate the values of all variables for the entire population (respondents and nonrespondents, whether originally sampled or not). The released data are samples from this synthetic population. The term "partially synthetic data" is now generally understood to mean that the released data contains a mixture of actual responses and simulated responses. All synthetic data applications combine sophisticated multivariate statistical modeling and computationally intensive simulations based on this modeling. Abowd and Woodcock (2001, 2004) showed that the partially synthetic approach could be reliably used to produce releaseable microdata files based on confidential data integrated from employer, household, and job-level records.

In February 2006, the Census Bureau released a pilot web-based application known as "On The Map"[1]. This mapping tool allows the user to select an arbitrary geographical region (resolved to the Census block) to show where people live and work. Companion reports on the workers' age, earnings, and industry distributions, as well as the number of employers, and Quarterly Workforce Indicators (QWI)[2], are resolved to the block group level. On The Map is the first synthetic data product publicly released by the Census Bureau. The QWI (released initially in 2003) were the Census Bureau's first public use product protected by noise infusion, a somewhat simpler version of the same principles used in synthetic data.

Although On The Map was released first, the Longitudinal Employer-Household Dynamics (LEHD) Program has been engaged in synthetic data research since 2001. The most ambitious project, which is nearing completion, has been to create a partially synthetic version of the 1990-1996 panels of the Survey of Income and Program Participation (SIPP) that have been linked to Social Security Administration (SSA) benefits data and Internal Revenue Service (IRS) data from the complete longitudinal earnings history of the respondent (back to 1950). The design of this file is the product of collaboration between the Census Bureau; SSA/Office of Research,

[1] Available at http://lehd.dsd.census.gov on September 10, 2006.

[2] These are flagship products for the LEHD program providing unprecedented information on employment, churning, and earnings that describe the local employment dynamics. For additional details, visit http://lehd.dsd.census.gov/led/library/tech_user_guides.html, available on September 10, 2006.

Evaluation, and Statistics; IRS/Statistics of Income and Research; and the Congressional Budget Office. Because the existing versions of the SIPP public use files already include thousands of variables on each respondent household, the version under development for this project synthesizes every variable except for a small group of demographic and benefit variables (four variables in all). The remaining 500+ variables on the file are synthetic.

Currently the SIPP is undergoing re-engineering to become DEWS. The knowledge gained from the project described above has been used by the re-engineering committee to help assess the costs and benefits of including more data from administrative sources on future SIPP public use releases.

TECHNICAL BACKGROUND

In developing its public use products based on synthetic data, LEHD employs a pragmatic approach which Little (2006) has termed "calibrated Bayes." Our public-use synthetic data products are based on estimated posterior predictive distributions. The prior contribution is based on "parameters" that are a combination of empirical estimates and constants that are controlled for confidentiality protection. The likelihood contribution is based on the actual confidential data, as well as estimated using a variety of techniques including exact multivariate distributions and approximations. The posterior predictive distributions are used to make multiple draws, known as implicates, by high-power computers to produce the synthetic data.

LEHD integrates existing data from administrative records, censuses, and surveys to build its infrastructure file system. Multiple imputation is used to replace missing values, and noise infusion is applied strategically to provide confidentiality protection for public use statistics based on establishment-level summaries such as QWI. The LEHD program uses only a small amount of cell suppression and is migrating away from this technique by implementing synthetic data replacements for the suppressed values.

The goal of LEHD is to build a longitudinal national frame of jobs[3] with an associated data infrastructure to support rapid production of results. The present infrastructure includes the history and characteristics for each worker and each employer under the state wage record system.

Since a worker may have multiple jobs, synthetic observations in On The Map are generated for both the count of jobs and the count of workers for each unique census block of residence (also called origin), conditional on each census block of workplace (also called destination) and defined combination of age,[4] earnings,[5] and industry[6] association (collectively called characteristics). Minnesota is the only state in the U.S. that codes the worker's establishment on

[3] Currently based on state unemployment wage records and the Quarterly Census of Employment and Wage (QCEW) records that cover about 98 percent all private, non-farm employment.

[4] There are three categories for the age of the worker – up to 30, 31-54, and 55 and above.

[5] There are three categories for the average monthly earnings of the highest paying job of a worker in a quarter – up to $1,199, $1,200 – $3,399, and $3,400 and above.

[6] There are 20 categories based on the 2-digit sector level under the North American Industry Classification (NAICS) system, available at http://www.census.gov/epcd/www/naics.html on September 10, 2006.

its unemployment insurance wage records. Consequently, the Minnesota wage record data can be directly integrated with the establishment level data using conventional, identifier-based record linkage methods. For all other states, LEHD uses a statistical model, estimated using information from Minnesota and information from the specific state, to multiply impute the establishment (and therefore destination) for each worker employed in a multi-unit firm.

The synthetic data for On The Map are generated for each origin based on posterior Dirichlet (multivariate beta) distributions conditional on unique destination and characteristics. The conjugate prior distribution is also a conditional Dirichlet distribution which must have sufficient empirical support and whose parameters are specified for confidentiality protection. The likelihood function is the multinomial distribution conditional on destination and characteristics. Noise is already infused in the destination count of workers. Thus, only the origin data are synthesized, and On The Map may be described as a partially synthetic data product, using the current standard nomenclature. Ten implicates were drawn from the posterior predictive distribution for each origin-destination pair, and the first implicate was used for the implementation of On The Map.

The SIPP/SSA/IRS Public Use File (PUF) and the proposed administrative records enhancements to DEWS are based on a more elaborate synthetic data system that works variable by variable to create the synthetic values. The system is based on sequential regression multivariate imputation (SRMI, Raghunathan, Reiter and Rubin 2003) as applied to the partially synthetic data with missing values (Reiter 2004). The estimation system stratifies the analysis based on non-synthesized variables and other variables specified by the analyst building the synthetic file. The imputation/synthesizing software automatically builds a posterior predictive distribution for each variable, conditional on all other values in the data. First, the missing data are completed using SRMI. Then the synthetic implicates are built using the same software.

The value of synthetic data lies in its ability to provide public use microdata that preserve analytical validity and provide confidentiality protection.

EVALUATION CRITERIA

As a state-of-the-art innovative approach, integrated data and synthetic data do not have comparable quality standards as in census and survey data. This situation is perhaps similar to the period when the concept of random sampling was introduced at the end of the nineteenth century (Wu 1995), but contemporary statistical theories and methods to support random sampling did not get fully established in the international statistical community until more than 30 years later.

Development of LEHD products follows the general guidelines of the Data Quality Act (2001) and the Census Bureau (2006) on utility, objectivity, and integrity. In particular, LEHD products have been developed to fit the needs of its users.

The creation of a voluntary federal-state partnership known as Local Employment Dynamics (LED) was based on the recognition and principle that the state partners supply their wage records on workers and firms and the Census Bureau build an integrated data infrastructure to create unprecedented new information about the local employment dynamics. Since the beginning of 4 states in 2000, LED has now grown to 43 state partners. The state of New York recently enacted legislation[7] to allow for data sharing with the Census Bureau in order to join LED. The Employment and Training Administration (ETA) of the U.S. Department of Labor recently agreed to provide partial funding support and pursue jointly with the Census Bureau to expand LED to a national program. The Brookings Institution identifies LEHD as a top priority for its federal data agenda[8].

LEHD was designed to use only existing data that have already been collected; therefore, it does not impose additional burden on the respondent. With LEHD as a highly automated operation, the cost of data processing is less than 2 cents per record although, in over-simplified terms, the annual LEHD volume is more than 20 times that of the decennial census[9] over a 10-year period.

On The Map was publicly released in 2006 after 18 months of intensive design, development, test, and evaluation guided by teams and beta testers inside and outside the Census Bureau. It was reviewed and approved by the Census Bureau Disclosure Review Board and verified for Section 508 compliance prior to its release.

The measures of analytic validity for On The Map are based on comparing the synthetic commuting distances with the distances computed from the underlying confidential data. Abowd, Andersson, and Roemer (2006) provide detailed evidence of the analytic validity of On The Map data.

Confidentiality protection is assessed using a measure that is comparable to the “swap rate” in systems that are based on data swapping. We compare the relative difference between the synthetic count and the actual confidential count of jobs or workers for all origin-destination pairs and their aggregates. Abowd et al (2006) introduced this “Reclassification Index,” which varies from 0 to 1. If the counts in synthetic and confidential data were identical in all cases, the reclassification index would equal to 0. The index may also be interpreted as the proportion of workers that need to be reallocated across origins in the synthetic data in order to replicate the actual data. The analysis of the reclassification index shows that for small geographic areas, there is considerable reclassification required to reconstruct the confidential data (often more than 50% of the cases must be reallocated) whereas in large geographic area only a trivial percentage need reclassification (usually less than 2%).

Since its release, On The Map has stimulated strong interests and discussions not only about its use for workforce and economic development, but also for transportation planning, emergency preparedness and response, and military base realignment. Originally funded for 12 pilot states by ETA, there are now 16 states participating in On The Map, and ETA has agreed to support the Census Bureau to expand the application to as many as 44 states in 2007.

[7] New York State Assembly bill is A11619 and New York State Senate bill is S08072, available at http://assembly.state.ny.us/leg/ on September 10, 2006.

[8] Available at http://www.brookings.edu/metro/umi/lehd.htm on September 10, 2006.

[9] There will be approximately 300 million individual records for the 2010 census, which is conducted every 10 years. There are about 150 million workers in the nation whose records are processed four times a year by LEHD.

On The Map public use data are currently made available through its state partners, the Census Bureau, and the Cornell University. In particular, all 10 implicates of On The Map data for Oregon and Texas are currently available for use and evaluation[10] by registered users. Users are also encouraged to submit emails[11] to comment on the program and report bugs[12] in the application, as well as share their findings and results through listserves.

On The Map is a relative simple application of the synthetic data approach. For the SIPP/SSA/IRS PUF and DEWS applications both the analytic validity studies and the confidentiality protection analysis are much more complicated. Analytic validity is assessed by comparing the complete univariate distribution of each synthetic variable to its confidential counterpart. Discrete variables match exactly for the overall sample and the sub-samples represented by the unsynthesized variables. Continuous variables match every percentile from 1 to 99 exactly. Multivariate analytic validity is assessed using covariance matrices, regression analyses, micro-simulation, and propensity score methods.

Confidentiality protection in these more complicated partially synthetic applications is based on attempting to re-identify the confidential source record for each synthetic record. Probabilistic record linking and distance matching are used for the re-identification studies. The goal is to have very low overall correct re-identification rates and to have the "best match" case be a false re-identification as often as it is a true re-identification. Both of these standards have been met.

ONGOING DEVELOPMENT

Given its brief history, development and practical use of synthetic data is just beginning. The supporting theories and the measures for evaluating its usefulness and value will undoubtedly grow and evolve.

Through a 3-year grant to Cornell University as a coordinating institution with the Census Bureau as the prime subcontractor, the National Science Foundation (SES #0427889) encourages innovative, high-payoff research and education to develop public use synthetic data under the Census Bureau Research Data Center system, and to help facilitate collaboration to help design and test these products. LEHD benefits from and contributes to this activity.

In ensuring the maximum coverage of jobs, LEHD will pursue to establish a national program in the next two years. We are also acquiring and adding data sources to the job frame and infrastructure, including federal employment, postal employment, and non-employers such as the self-employed.

[10] Available at http://vrdc.ciser.cornell.edu/onthemap/doc/index.html on September 10, 2006.

[11] The standard email address is dsd.local.employment.dynamics@census.gov, and it is monitored directly by the program manager.

[12] Known and unresolved bugs about On The Map are posted at http://lehd.dsd.census.gov/led/datatools/onthemap.html, available on September 10, 2006.

The broad and growing interests about integrated data and synthetic data offer several business opportunities for the Census Bureau, covering workforce development, transportation planning, economic dislocation and development, emergency preparedness and response, in addition to aging research, small business policy analysis, and program evaluation. LEHD is also identified in the Census Bureau Strategic Plan[13] as a performance measurement tool on improvements to data coding, processing, and analysis. The transportation community, under the leadership of the American Association of State Highway and Transportation Officials and the US Department of Transportation (DOT), have also planned to compare LEHD data with the American Community Survey, the decennial census, and commercial data sources as part of its development of the next Census Transportation Planning Package. The ability to make longitudinal checks and edits of time series data is one of the inherent benefits of LEHD.

A particularly noteworthy outcome since the release of On The Map is the developing collaboration of the state labor market information (LMI) offices, the state DOT, and the local metropolitan planning organizations (MPO). As the state DOT and MPO have a growing desire for the origin-destination data, they also have better local knowledge about place of work. These agencies for at least 4 states have begun exchange of ideas to form a feed-forward and feedback loop to enhance data quality. This growing trend can benefit LEHD substantially by reducing the need for imputation and improving the overall data quality.

REFERENCES

Abowd, J.M, Andersson, F, and Roemer, M.I. (2006). The Disclosure Limitation Protocol for the Census Bureau's LEHD Prototype Transportation Package (draft).

Abowd, J.M. and S. Woodcock (2004). "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data," in J. Domingo-Ferrer and V. Torra (eds.) Privacy in Statistical Databases (New York: Springer-Verlag), pp. 290-297.

Abowd, J.M. and S. Woodcock (2001). "Disclosure Limitation in Longitudinal Linked Data," in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), (Amsterdam: North Holland), 215-277.

Census Bureau (2006). Census Bureau Section 515 Information Quality Guidelines. Available at http://www.census.gov/quality/ on September 10, 2006.

Duncan, G.T., de Wolf, V.A., Jabine, T.B., and Straf, M.L. (1993). Report of the Panel on Confidentiality and Data Access. Journal of Official Statistics, 9, 271-274.

Little, R (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9, 407-426.

[13] Objective 1.4: Produce new information using existing data sources by developing cutting-edge techniques and promoting knowledge sharing, available at http://www.census.gov/main/www/strategicplan/strategicplan.html#1-4 on September 10, 2006.

Little, R (2006). Calibrated Bayes: A Bayes/Frequentist Roadmap. The American Statistician. 60, 213-223.

Office of Management and Budget (2001). Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies. Available at http://www.whitehouse.gov/omb/fedreg/final_information_quality_guidelines.html on September 10, 2006.

Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, 19, 1-16.

Reiter, J.P. (2004). Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. Survey Methodology, 30, 235-242.

Rubin, D.B. (1993). Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, 9, 461-468.

Wu, J. (1995). One Hundred Years of Sampling, special invited paper in “Sampling Theory and Practice.” State Statistical Bureau of China, ISBN 7-5037-1670-3, China Statistical Publisher, Beijing, China.